ABSTRACT OF THE DISCLOSURE

Techniques for training and using linked event detection systems and transforming source-identified stopwords are provided. A training corpus of source identified stories and a reference language is determined. Optionally, stopwords for source-identified stories are transformed based on statistical analysis of parallel verified and un-verified transformations. Reference language and non-reference language terms are selectively included in source-pair term frequency-inverse story frequency models. Optionally, incremental source-identified term frequency-inverse story frequency models are determined. Selected terms are weighted and similarity metrics determined. Associated source-pair statistics, computed in part from a training corpus, are combined with the values of each similarity metric in the set of similarity metrics to form a similarity vector. Similarity vectors and verified link label information are used to determine a predictive model. Similarity vectors for story pairs are used with the predictive model to determine if the story-pairs are linked. Sources are arranged based on source inter-relationships into a source-hierarchy. Progressively more refined source-pair similarity statistics are also provided. New sources and associated source-pair similarity statistics are added by substituting related source-pair similarity statistics based on the source hierarchy and source characteristics. The source-pair similarity statistics are used to optionally normalize the similarity metrics.